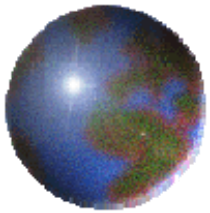


15. Descriptive Summary, Design, and Inference

Geographic Information Systems and Science

SECOND EDITION

Paul A. Longley, Michael F. Goodchild, David J. Maguire, David W. Rhind

Outline

- Data mining
- Descriptive summaries
- Optimization
- Hypothesis testing

Data mining

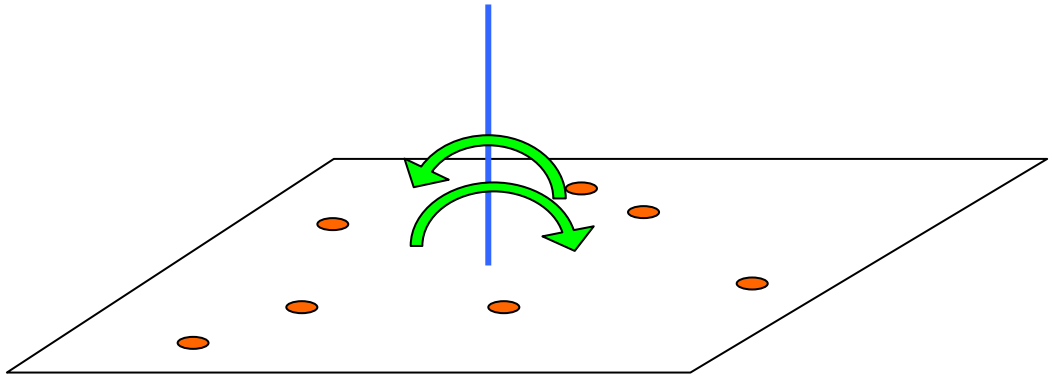
- Analysis of massive data sets in search for patterns, anomalies, and trends
  - spatial analysis applied on a large scale
  - must be semi-automated because of data volumes
  - widely used in practice, e.g. to detect unusual patterns in credit card use

Descriptive summaries

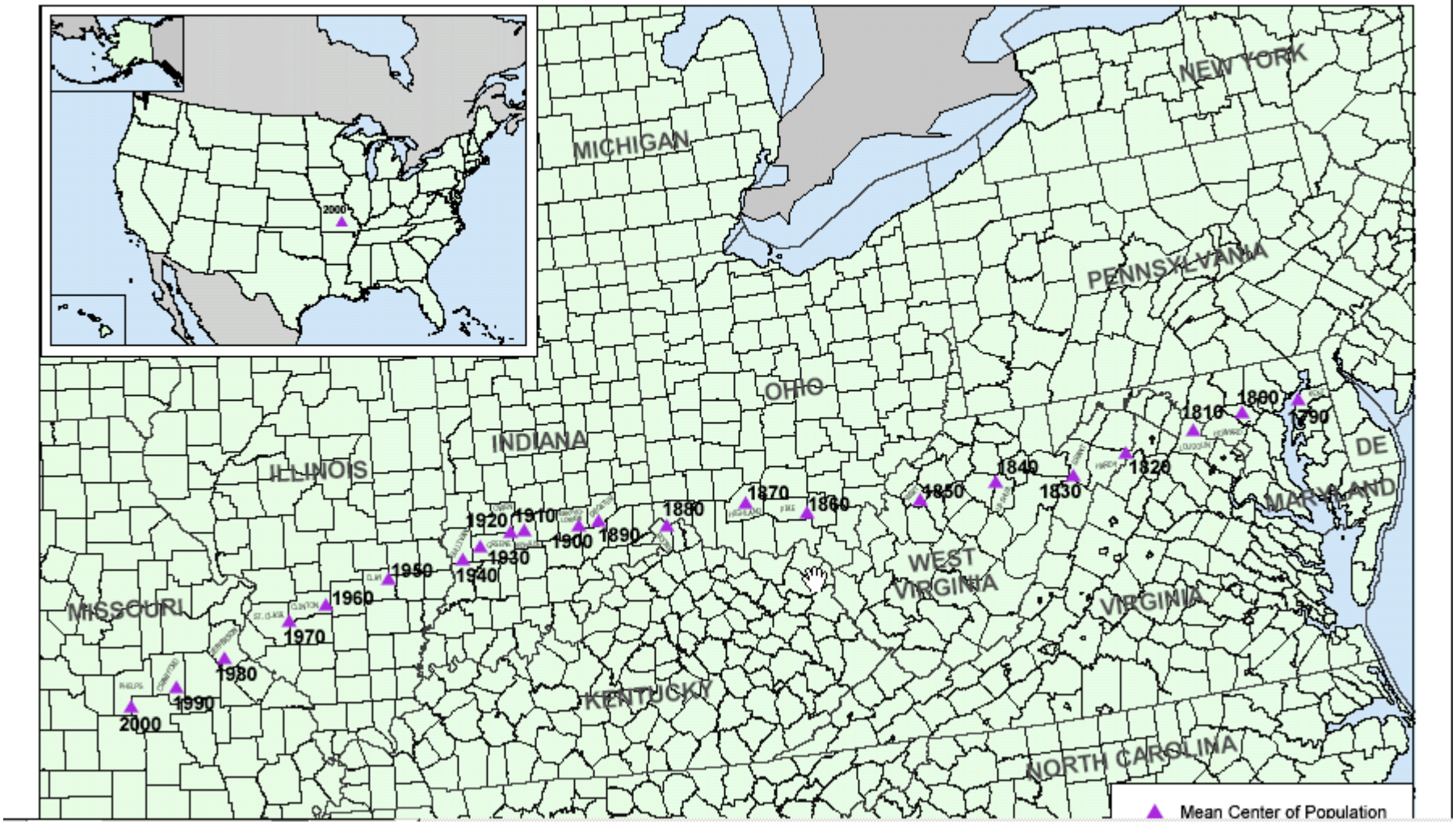
- Attempt to summarize useful properties of data sets in one or two statistics
- The mean or average is widely used to summarize data
  - centers are the spatial equivalent
  - there are several ways of defining centers

The centroid

- Found for a point set by taking the weighted average of coordinates
- The balance point



Mean Center of Population for the United States: 1790 to 2000



Optimization properties

- The centroid minimizes the sum of distances squared
  - but not the sum of distances from each point
  - the center with that property is called the point of minimum aggregate travel (MAT)
  - the properties have frequently been confused, e.g. by the U.S. Bureau of the Census in calculating the center of U.S. population
  - the MAT must be found by iteration rather than by calculation

Applications of the MAT

- Because it minimizes distance the MAT is a useful point at which to locate any central service
  - e.g., a school, hospital, store, fire station
  - finding the MAT is a simple instance of using spatial analysis for optimization

Dispersion

- A measure of the spread of points around a center
- Useful for determining positional error
- Related to the width of the kernel used in density estimation

Spatial dependence

- There are many ways of measuring this very important summary property
- The semivariogram, see Chapter 13
  - measures spatial dependence over a range of scales
- The Moran and Geary indices, see Chapter 5

Descriptions of Pattern

- Many techniques
  - depending on the type of features and whether they are differentiated by attributes (*labeled*)
    - measures for unlabeled features look for purely geometric pattern
    - measures for labeled features ask about patterns in the labels

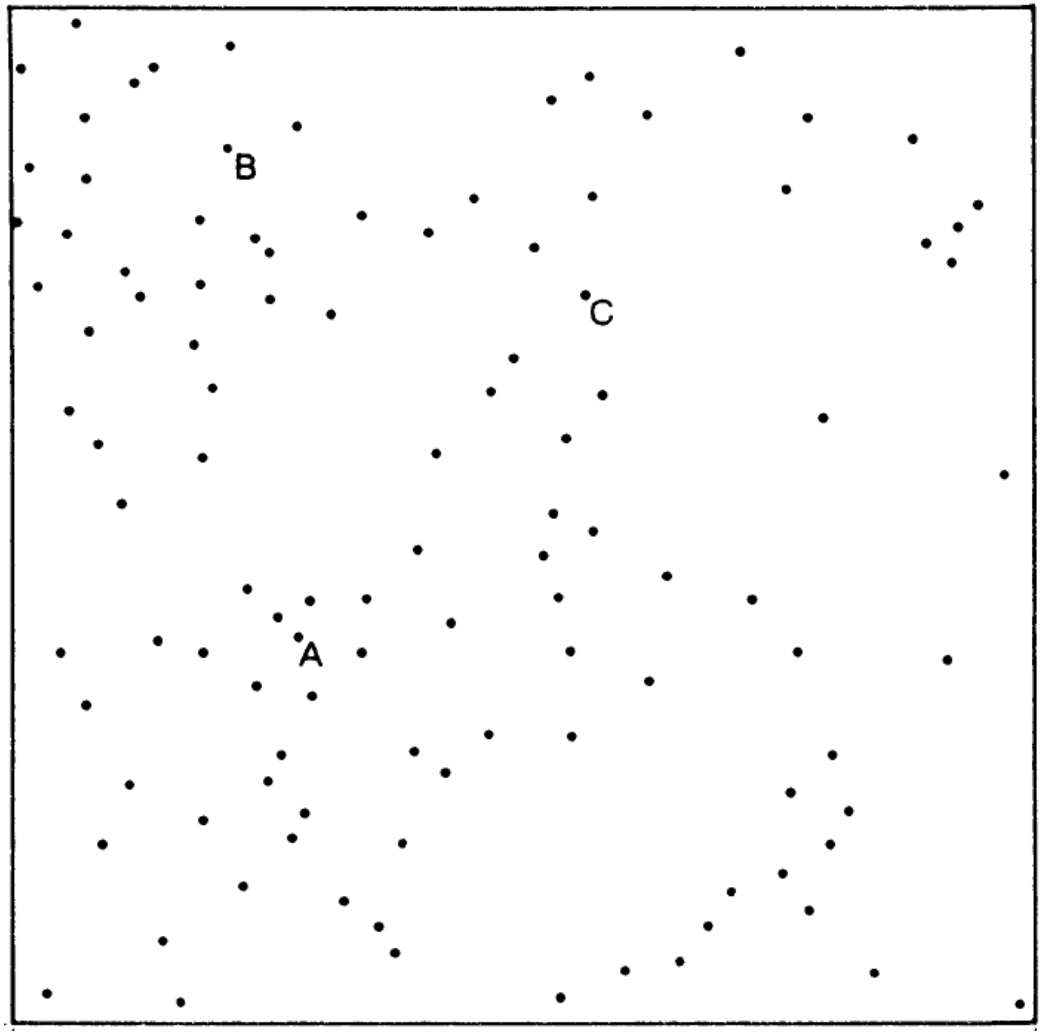
Patterns in Unlabeled Points

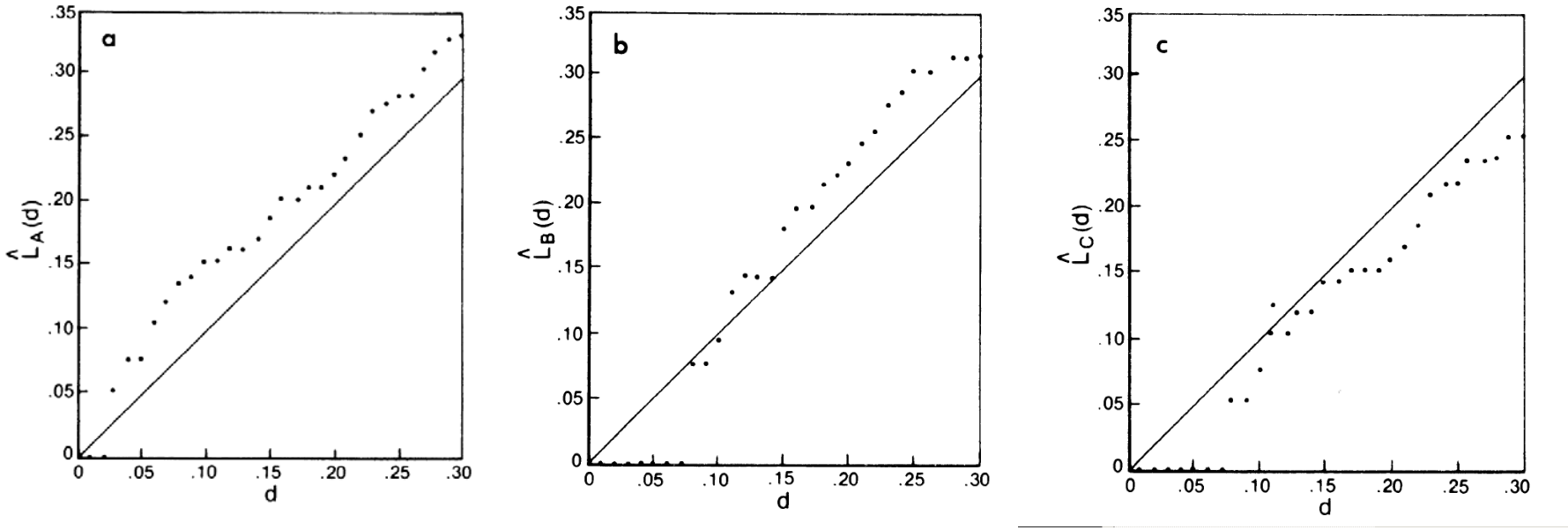
- Locations of disease, crimes, traffic accidents
  - Do events tend to *cluster* more in some areas than others?
  - Or are they *random*, equally likely anywhere?
  - Or are they *dispersed*, such that points are less likely in areas close to other points?

The K Function

- Captures how density of points varies with distance away from a reference point
  - By comparing to what would be expected in a random distribution of points

Point pattern of individual tree locations. A, B, and C identify the individual trees analyzed in the next slide. (Source: Getis A, Franklin J 1987 Second-order neighborhood analysis of mapped point patterns. *Ecology* 68(3): 473-477).



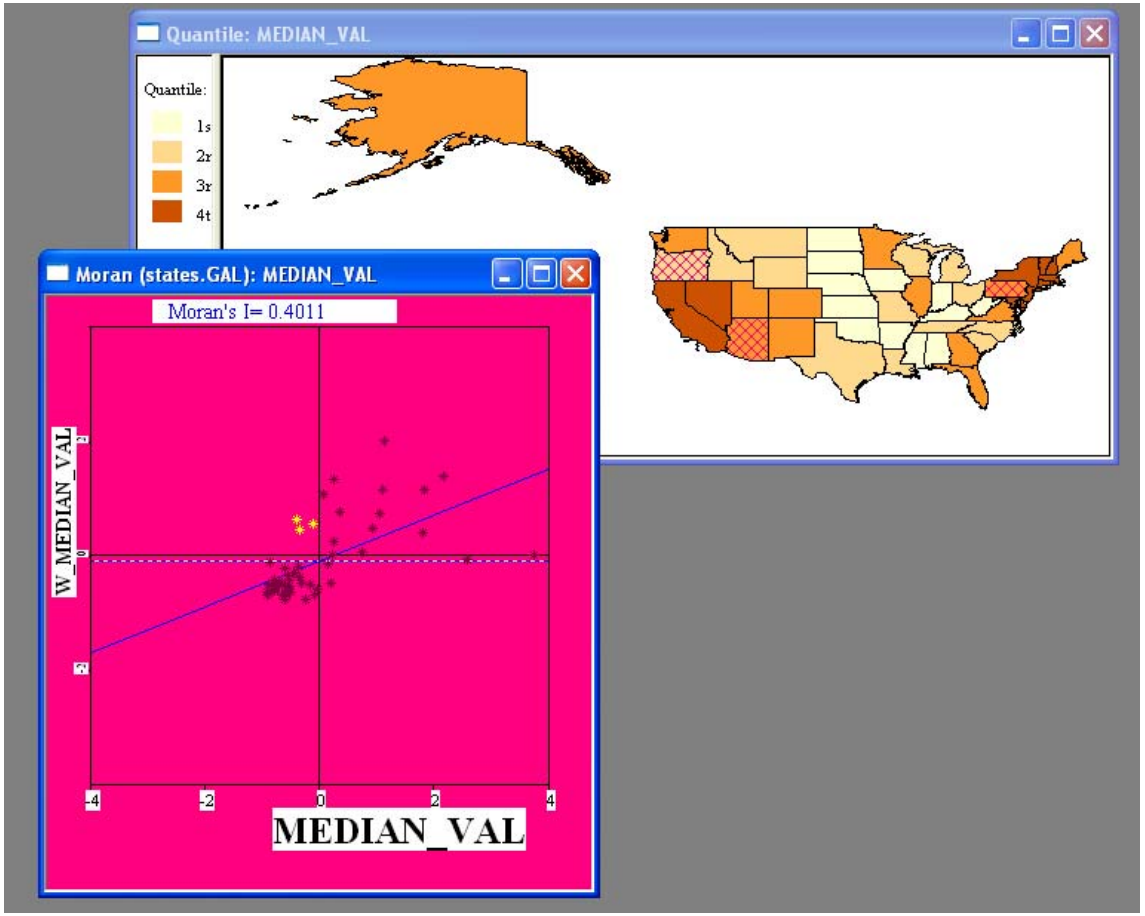


Analysis of the local distribution of trees around three reference trees in the previous slide (see text for discussion). (Source: Getis A, Franklin J 1987 Second-order neighborhood analysis of mapped point patterns. *Ecology* 68(3): 473-477).

Pattern in Labeled Features

- How are the attributes (labels) distributed over the features?
  - *Clustered*, with neighboring features having similar values
  - *Random*, with labels assigned independently of location
  - *Dispersed*, with neighboring features having dissimilar values

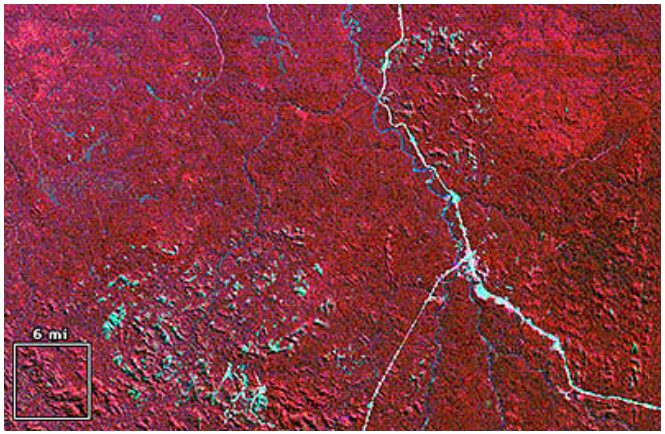
In the map window the states are colored according to median house value, with the darker shades corresponding to more expensive housing.

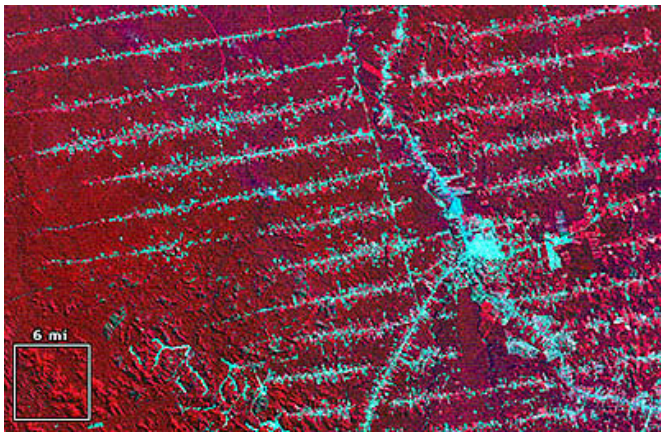


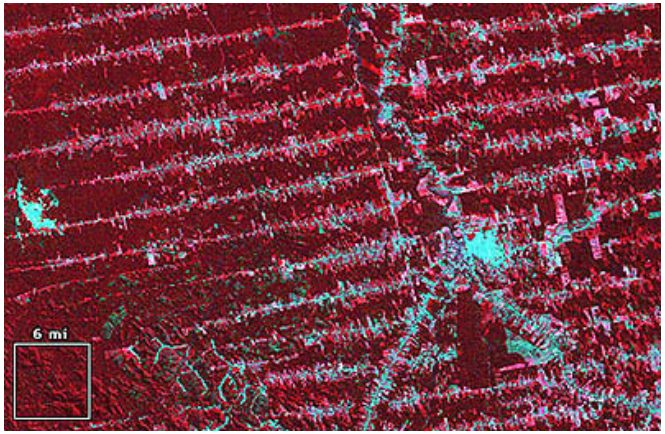
In the scatterplot window the three points colored yellow are instances where a state of below-average housing value is surrounded by states of above-average value.

Fragmentation statistics

- Measure the patchiness of data sets
  - e.g., of vegetation cover in an area
- Useful in landscape ecology, because of the importance of habitat fragmentation in determining the success of animal and bird populations
  - populations are less likely to survive in highly fragmented landscapes







Three images of part of the state of Rondonia in Brazil, for 1975, 1986, and 1992. Note the increasing fragmentation of the natural habitat as a result of settlement. Such fragmentation can adversely affect the success of wildlife populations.

Optimization

- Spatial analysis can be used to solve many problems of design
- A spatial decision support system (SDSS) is an adaptation of GIS aimed at solving a particular design problem

Optimizing point locations

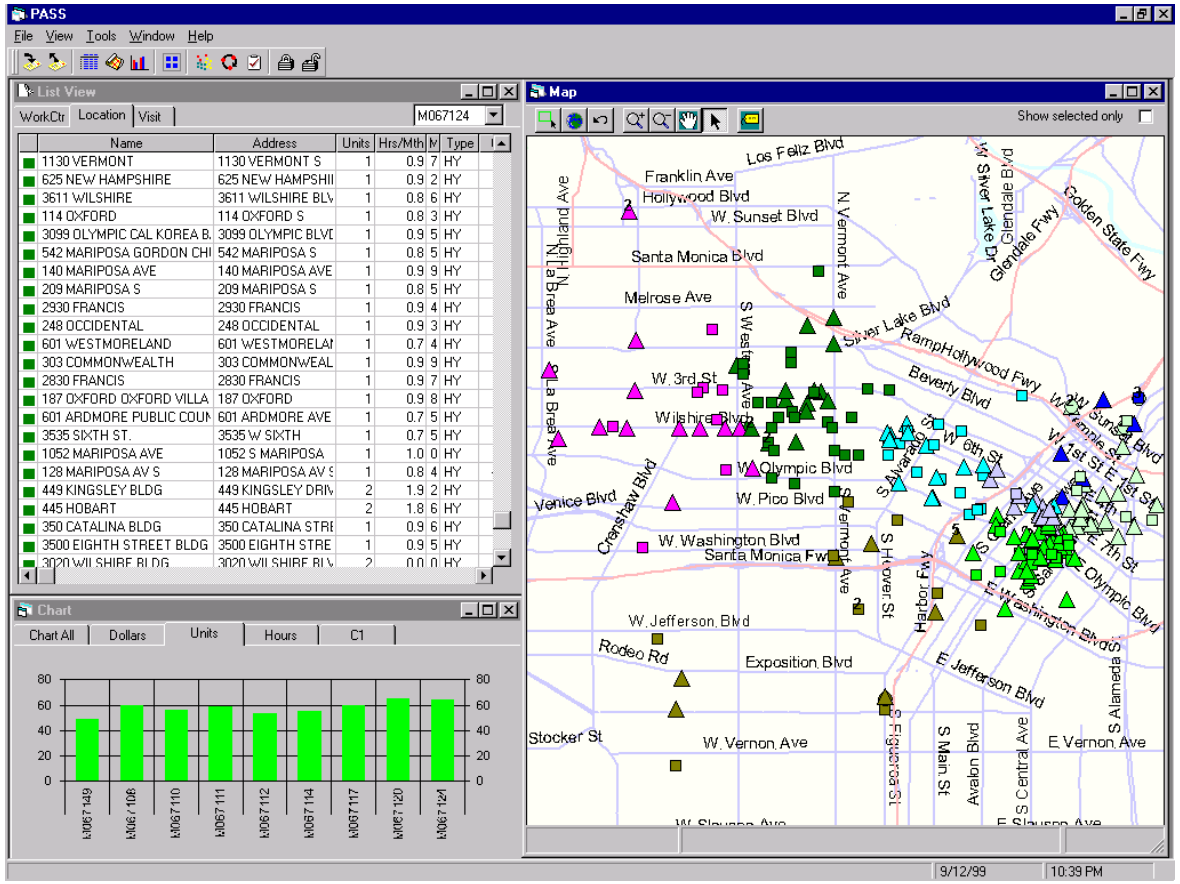
- The MAT is a simple case: one service location and the goal of minimizing total distance traveled
- The operator of a chain of convenience stores or fire stations might want to solve for many locations at once
  - where are the best locations to add new services?
  - which existing services should be dropped?

Location-allocation problems

- Design locations for services, and allocate demand to them, to achieve specified goals
- Goals might include:
  - minimizing total distance traveled
  - minimizing the largest distance traveled by any customer
  - maximizing profit
  - minimizing a combination of travel distance and facility operating cost

Routing problems

- Search for optimum routes among several destinations
- The traveling salesman problem
  - find the shortest tour from an origin, through a set of destinations, and back to the origin



Routing service technicians for Schindler Elevator. Every day this company's service crews must visit a different set of locations in Los Angeles. GIS is used to partition the day's workload among the crews and trucks (color coding) and to optimize the route to minimize time and cost.

Optimum paths

- Find the best path across a continuous cost surface
  - between defined origin and destination
  - to minimize total cost
  - cost may combine construction, environmental impact, land acquisition, and operating cost
  - used to locate highways, power lines, pipelines
  - requires a raster representation

Solution of a least-cost path problem. The white line represents the optimum solution, or path of least total cost, across a friction surface represented as a raster. The area is dominated by a mountain range, and cost is determined by elevation and slope. The best route uses a narrow pass through the range. The blue line results from solving the same problem using a coarser raster.

Hypothesis testing

- Hypothesis testing is a recognized branch of statistics
- A sample is analyzed, and inferences are made about the population from which the sample was drawn
- The sample must normally be drawn randomly and independently from the population

Hypothesis testing with spatial data

- Frequently the data represent all that are available
  - e.g., all of the census tracts of Los Angeles
- It is consequently difficult to think of such data as a random sample of anything
  - not a random sample of all census tracts
- Tobler's Law guarantees that independence is problematic
  - unless samples are drawn very far apart

Possible approaches to inference

- Treat the data as one of a very large number of possible spatial arrangements
  - useful for testing for significant spatial patterns
- Discard data until cases are independent
  - no one likes to discard data
- Use models that account directly for spatial dependence
- Be content with descriptions and avoid inference